Review of Master's Thesis

Student: Kovařík David, Bc.
Title: Automatic Template Pattern Recognition (id 20198)
Reviewer: Lengál Ondřej, Ing., Ph.D., UITS FIT VUT

1. Assignment complexity — **average assignment**

2. Completeness of assignment requirements — **assignment fulfilled**

3. Length of technical report — **in usual extent**

4. Presentation level of technical report — **90 p. (A)**
Prezentační úroveň technické zprávy je dobrá. Popisy algoritmů jsou vhodně doplněny příklady.

5. Formal aspects of technical report — **85 p. (B)**
Práce je psána nadprůměrně dobrou angličtinou. V popisech algoritmů bych ocenil větší formálnost (vstupy, výstupy, atp.), občas jsem měl problém pochopit, co některé příkazy dělají, podobně i v textu. Typografie je v pořádku. Text v obrázcích by mohl být větší a lépe čitelný.

6. Literature usage — **75 p. (C)**
Student využil relevantní zdroje, chybí mi však vymezení se vůči technikám detekce spamu v emailu a také vůči technikám učení konečných automatů založených na algoritmu L*. Jako jedna z optimalizací výstupu je prezentována optimalizace velikosti regulárních výrazů, kde podle mě již existují řešení (například minimalizace konečného automatu pro výraz a následná transformace do reg. výrazu).

7. Implementation results — **85 p. (B)**
Student vytvořil prototypovou implementaci v jazyce Python, která se ze vstupní spamové kampaně učí její model. Výstupem programu je daný model ve formě regulárního výrazu. Uživatel si též může zobrazit výsledek klasifikace množiny SMS zpráv podle výstupního modelu. Kód je přehledný a vhodně komentovaný.

8. Utilizability of results
Práce je zaměřena na učení modelů pro popis množiny SMS zpráv charakterizující spamovou kampaň. Navrhuje nové řešení založené na algoritmech pro zarovnávání řetězců. Výstup práce může sloužit jako základ pro vývoj systému detekce spamu v SMS zprávách mobilními operátory.

9. Questions for defence
1. Proč nelze při učení použít negativní protipříklady, například zprávy z běžné SMS komunikace?

2. V práci zmiňujete jako nevýhodu některých optimalizací to, že je nutné tokenizovat i zprávy, které by jinak tokenizovány nebyly, čímž trpí výkon. Je to opravdu releventní? Tokenizování by měla být lineární operace (zpráva je stejně testována na příslušnost do jazyka regulárního výrazu, což je velmi podobná operace). Jak velké zpomalení způsobí tato tokenizace?

10. Total assessment — **84 p. very good (B)**
Práce pana Kovaříka se zabývá syntézou modelů popisujících spamové SMS kampaně. Cílem práce bylo vyvinout systém, který se z kolekce SMS zpráv patřících do SMS spam kampaně naučí šablonu, která tuto kampaň popisuje. Student navrhl algoritmus založený na zarovnávání zpráv a zobecňování na základě lišících se podřetězců. Algoritmus je implementován v prototypu doplněném o HTML výstup a jednoduché uživatelské rozhraní.

Práce řeší (dle mého názoru) zajímavé téma přístupem, který pro mnoho instancích v uvažované testovací sadě funguje dobře (existují ale i sady zpráv, na kterých algoritmus selže). **Práci hodnotím celkově jako kvalitní**, za její silné stránky považuji návrh různých heuristiky, jejichž potřeba vyvstala až v průběhu práce, a implementace celého rámce do nástroje umožňující uživateli práci využít. Za slabší stránku považuji mezery v použité literatuře a slabší formální popis použitých metod. Celkově hodnotím práci pana Kovaříka **stupněm B**.

In Brno 1. June 2018

.....
signature